Scientific and Technical Report

Sponsored by
Advanced Research Projects Agency/ITO
and United States Patent and Trademark Office

Browsing, Discovery and Search in Large Distributed Databases
of Complex and Scanned Documents

ARPA Order No. D570

Issued by EXC/AXS under Contract #F19628-95-C-0235

Date Submitted: October 5, 1999

Period of Report: July 1, 1999 to September 30, 1999

Submitted by: Professor W. Bruce Croft, Principal Investigator
Computer Science Department
University of Massachusetts, Amherst

Distribution Statement A: Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	10/05/99	Scientific/Tech 7/1/99 - 9/30/99

4. TITLE AND SUBTITLE
Browsing, Discovery, and Search in Large Distributed Databases of Complex and Scanned Documents

5. FUNDING NUMBERS
F19628-95-C-0235
ARPA Order No. D570

6. AUTHOR(S)
W. Bruce Croft

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts, Amherst
Box 36010, OGCA, Munson Hall
Amherst, MA 01003-6010

8. PERFORMING ORGANIZATION REPORT NUMBER
TR5281811099

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Mr. Charles Shank
ESC/PKRB
104 Barksdale St., Bldg 1520
Hanscom AFB, MA 01731-1806

Ms. Monique Dillon
Office of Naval Research
Boston Regional Office
495 Summer St., Room 103
Boston, MA 02210-2109

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This project aims to integrate powerful, new techniques for interactive browsing, discovery, and retrieval in very large, distributed databases of complex and scanned documents. Emphasis is placed on going beyond full-text retrieval techniques developed in the DARPA TIPSTER program to support different types of access and non-textual content. These techniques should be particularly relevant to the patent domain where it is important to find relationships between documents and where the patent or trademark may be based on a visual design. The specific tasks identified involve studying representation techniques for long documents with complex structure, browsing and discovery techniques for large text databases, image retrieval and scanned document retrieval techniques, and architectures for large, distributed databases.

14. SUBJECT TERMS			15. NUMBER OF PAGES
Browsing	Query Processing	Indexing	10
Image Retrieval	Scanned Document Retrieval	Bayesian Network	16. PRICE CODE
Text Retrieval	Probabilistic Retrieval Model	Large Distributed Databases	

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

PTO-Related Bibliography

MM-18: (1997) Wu., V., Manmatha, R. and Riseman, E. "TextFinder: An Automatic System to Detect and Recognize Text in Images," to appear in *Transactions on Pattern Analysis & Machine Intelligence,"* CIIR technical report.

IR-152: (1998) Leouski, A., and Allan, "Strategy-based Interactive Cluster Visualization for Information Retrieval," to appear in *The International Journal on Digital Libraries.*

IR-180: (1999) Callan, J. and Connell, M., "Query-Based Sampling of Text Databases," submitted to ACM TOIS.

Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents

Technical and Scientific Report

Task 1: Representation Techniques for Complex Documents

Task Objectives

In this task, the goal is to extend the word-based representations that are common in retrieval systems in order to support summarization, browsing, and more effective retrieval. Specifically, we have been studying phrase-based representations and relationships between phrases in individual and groups of documents as the basis for our approach. Document structure will be used as part of the information that is used to "tag" the phrasal representation.

Technical Problems

The technical problems have to do with defining a "phrase", developing techniques for rapidly extracting them from text, comparing phrase contexts to identify significant relationships, producing summaries from these representations, extending the underlying retrieval model to be able to make effective use of phrasal representations, and using complex document structure in indexing and retrieval.

General Methodology

The general methodology for this task is to demonstrate effectiveness through user-based and collection-based experiments. As well as the PTO text databases, we will make extensive use of the TIPSTER document collection, which consists of a large number of text documents from a variety of sources, queries, and user relevance judgments for each query.

Technical Results

We are currently carrying out research on deciding when it is better to use a phrase rather than to use the single terms that make up the phrase in a search. Our research currently shows that when the single terms occur frequently, but the phrase occurs infrequently, it is better to use the phrase. We are still investigating the best way to incorporate these finding into our query formulation algorithms. We are also performing exploratory data analyses on the statistical distributions of phrases and the terms that make up the phrases, in whole collections, and in relevant documents.

We have developed a new retrieval technique based on language models. One advantage of language models is that they provide a better theoretical foundation for retrieval. This approach has showed a lot of promise with some data – in fact initial experiments with a crude version of the model have shown that it can perform as well as other systems at TREC. We continue to carry out experiments to see the effects of learned language models on retrieval.

Important Findings and Conclusions

None.

Significant Hardware Development

None.

Special Comments

None.

Implication for Further Research

Language models may provide retrieval improvements for PTO data.

It may be possible to develop algorithms to decide when phrases should be preferred over single terms and vice-versa.

Task 2: Browsing and Classification Techniques for Document Collections

Task Objectives

The goals of this task are to develop techniques for summarizing and classifying collections of documents. These techniques will be designed to support interactive browsing and text classification in environments like the PTO.

Technical Problems

The technical problems involve producing an effective summary of a group of documents, such as a retrieved set or an entire database. Both document and phrase clusters could be used as part of this process. The classification task emphasizes the ability to accurately assign predefined categories (as in the PTO classification) to new documents (patents). An additional problem is to determine when existing classifications do not match well to new documents, such as when a PTO category covers too many patents and needs to be refined.

General Methodology

Evaluation of these techniques will be done using both the TREC corpus and PTO data. For the classification task in particular, we are designing evaluation criteria with substantial input from PTO staff.

Technical Results

The results from distributed retrieval (see Task 4) show that database sampling is an effective method for finding resource descriptions. We are, therefore, working on a multi-level scheme for classification which involves dividing a database into smaller databases by class and then using collection selection to find the appropriate class. A number of different approaches to collection selection are being investigated.

We are currently carrying out a project with Dataware Technologies to evaluate different approaches to classification.

In the summarization/visualization area, we have developed a system for combining a ranked list with clustering. A ranked list is a well-known technique for presenting information so that relevant documents may be found quickly. Clustering is also a well-known technique for grouping similar documents. By combining the two, we have developed an approach that exceeds the retrieval effectiveness of a traditional ranked list. This work is described in the following paper:

- Leuski, A., and Allan, J., "Strategy based Interactive Cluster Visualization for Information Retrieval". To appear in The International Journal on Digital Libraries.

Important Findings and Conclusions

None.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

We will continue to improve the demonstration system and plan to carry out further classification experiments using the collection mechanism. We are evaluating different approaches to classification with Dataware.

Task 3: Image Indexing and Retrieval

Task Objectives

The goal of this task is to develop similarity-based techniques for retrieving images such as trademarks, logos, and designs.

Technical Problems

The central issue is how images can be indexed to support efficient, content-based retrieval. The primary type of query in these environments is "find me things that look like this". We are developing "appearance-based" retrieval of images as well as more straightforward features such as color and texture. Filter based and frequency domain based techniques offer some potential in this area, but significant work needs to be done on making this approach efficient enough to deal with hundreds of thousands of images.

General Methodology

The evaluation of these techniques will be done in a similar way to text by developing test collections of images. Specifically, we are working to obtain large collections of trademark and design images, both from the PTO and from general sources such as the web.

Technical Results

Our work is now focused on evaluating the demonstration system. We have received a complete collection of geometric trademarks with relevance judgements from the British Patent Office. The relevance judgements were performed by a trademark examiner. We are continuing to index this database so that the performance of the demonstration system can be evaluated. We have created an initial version of a user interface for obtaining relevance judgements on the USPTO database so that we may be able to evaluate the effectiveness of the trademark demonstration system on the PTO database. The next step will be to obtain relevance judgements on the USPTO database using this interface. We have also implemented a method to retrieve images using invariant moments which we would like to use for comparison (There are some researchers who have used moments to try to retrieve trademarks). We hope to show in our evaluations that our techniques work much better than invariant moments. We also continue to improve the effectiveness of our trademark retrieval system.

We have collected 4000 additional flower images from the web. We are currently indexing these images and adding them to the flower patent database to judge retrieval effectiveness over a larger set.

We have carried out initial work showing that our technique for segmenting flowers from images may also be used for segmenting other objects like birds. Much more work needs

to be done in this area, but the initial work shows this to be a promising approach for segmenting objects from (certain kinds of) images so that they can be further indexed for retrieval.

Important Findings and Conclusions

None.

Significant Hardware Development

None

Special Comments

None.

Implications for Further Research

We continue to focus on evaluating the accuracy of our techniques using trademark testbeds from Britain and, hopefully, from the U.S. PTO. We will also continue to improve the demonstration system.

Task 4: Distributed Retrieval Architecture

Task Objectives

The goals of this task are to scale up our current methods of automatically selecting collections and merging results, and to investigate architectures that can support efficient retrieval, browsing and relevance feedback in distributed environments with terabytes of information.

Technical Problems

The current INQUERY text retrieval system uses a client-server architecture to support simultaneous retrieval from multiple collections distributed across one or more processors. A number of efficiency bottlenecks develop, however, when the size of the databases is very large. Deciding which subcollections to search can address part of the problem, but there are other problems associated with the fundamental efficiency of the processes involved and the use of distributed resources. Image indexing and retrieval tends to make all of these problems worse since the databases and indexes are considerably larger.

General Methodology

The architectures and algorithms produced in this task will be evaluated using a combination of standard performance (efficiency) measures and effectiveness measures. The efficiency tests will be done using TREC data and large PTO databases, including images, and the collection selection algorithms will be evaluated using the text subcollections of the patents.

Technical Results

Completed experiments studying the accuracy of document retrieval using resource descriptions obtained by sampling remote resources. The results on a testbed containing 100 databases and 100 queries showed that there was no noticeable loss of effectiveness when using resource descriptions obtained by sampling. This result is important because database sampling is the first practical method of discovering (or verifying) the contents of databases controlled by third parties. This work is described in the following paper which has been submitted to the journal ACM TOIS:

- J. Callan and M. Connell. "Query-Based Sampling of Text Databases". Submitted to ACM Transactions on Information Processing.

Results from distributed retrieval are being used for the classification problem (see Task 2).

We are in the process of obtaining relevance judgements to determine whether it is better to organize documents chronologically or by subject. Results from other research suggest that it is better to organize by subject. We are verifying with PTO data. We are also examining four different algorithms for merging databases. This work is also being carried out using PTO data.

More experiments are being carried out with the language model approach to collection selection to evaluate its effectiveness in terms of precision/recall.

Important Findings and Conclusions

Organization by topic is likely to be better than a chronological organization. Distributed search can be more effective than centralized search if it is based on language models. Database sampling is the first practical method of discovering (or verifying) the contents of databases controlled by third parties.

Significant Hardware Development

None.

Special Comments

None

Implications for Further Research

Organizing documents by subject is likely to be more effective than organizing them by the date of the document. We will continue to evaluate performance of distributed architectures for scalable IR using the new demonstration system. Database sampling is a good practical way of discovering (or verifying) the contents of databases controlled by third parties.